




# Automated conversational agents for post-intervention follow-up: a systematic review

L. Geoghegan <sup>1,\*</sup>, A. Scarborough<sup>2</sup>, J. C. R. Wormald<sup>3</sup>, C. J. Harrison<sup>3</sup>, D. Collins<sup>4</sup>, M. Gardiner <sup>5</sup>, J. Bruce <sup>6</sup> and J. N. Rodrigues<sup>6,7</sup>

<sup>1</sup>Section of Vascular Surgery, Department of Surgery and Cancer, Imperial College London, London, UK

<sup>2</sup>Department of Cardiothoracic Surgery, King's College Hospital, London, UK

<sup>3</sup>Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

<sup>4</sup>Department of Plastic, Reconstructive and Burns Surgery, Chelsea and Westminster Hospital, London, UK

<sup>5</sup>Department of Plastic and Reconstructive Surgery, Frimley Park Hospital, Guildford, UK

<sup>6</sup>Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK

<sup>7</sup>Department of Plastic and Reconstructive Surgery, Stoke Mandeville Hospital, Aylesbury, UK

\*Correspondence to: Section of Vascular Surgery, Department of Surgery and Cancer, Academic Section of Vascular Surgery, Imperial College London, Exhibition Road, London SW7 2AZ, UK (e-mail: luke.geoghegan13@imperial.ac.uk)

## Abstract

**Background:** Advances in natural language processing and other machine learning techniques have led to the development of automated agents (chatbots) that mimic human conversation. These systems have mainly been used in commercial settings, and within medicine, for symptom checking and psychotherapy. The aim of this systematic review was to determine the acceptability and implementation success of chatbots in the follow-up of patients who have undergone a physical healthcare intervention.

**Methods:** A systematic review of MEDLINE, MEDLINE In-process, EMBASE, PsychINFO, CINAHL, CENTRAL and the grey literature using a PRISMA-compliant methodology up to September 2020 was conducted. Abstract screening and data extraction were performed in duplicate. Risk of bias and quality assessments were performed for each study.

**Results:** The search identified 904 studies of which 10 met full inclusion criteria: three randomised control trials, one non-randomised clinical trial and six cohort studies. Chatbots were used for monitoring after the management of cancer, hypertension and asthma, orthopaedic intervention, ureteroscopy and intervention for varicose veins. All chatbots were deployed on mobile devices. A number of metrics were identified and ranged from a 31 per cent chatbot engagement rate to a 97 per cent response rate for system-generated questions. No study examined patient safety.

**Conclusion:** A range of chatbot builds and uses was identified. Further investigation of acceptability, efficacy and mechanistic evaluation in outpatient care pathways may lend support to implementation in routine clinical care.

## Introduction

The first known agent capable of conversation between human and machine was developed in 1966<sup>1</sup>. Eliza used early natural language processing to return open-ended questions to users, simulating person-centred psychotherapy.

Developments in speech recognition, natural language processing, natural language understanding and artificial intelligence have led to the design of systems capable of mimicking human interaction with unconstrained natural language input<sup>2</sup>. A chatbot is defined as 'a computer program designed to simulate conversation with human users, particularly over the internet'<sup>3</sup>. A recent systematic review involving 17 studies and 1573 participants found that chatbots in healthcare were predominantly used in mental health conditions to educate patients and collect data from health-related questionnaires<sup>4</sup>.

Financial pressures and clinical demand have driven interest in virtual clinics for monitoring and surveillance following healthcare interventions<sup>5</sup>, particularly during the COVID-19 pandemic, with rapid adoption of virtual services to moderate

infection risk through reduction of direct clinician-patient contact<sup>6</sup>. A recent randomised trial involving 209 general surgical patients demonstrated better attendance (92 versus 81 per cent) and higher patient satisfaction (95 per cent of participants happy or very happy versus 56 per cent) with virtual postoperative clinics compared with traditional outpatient follow-up<sup>7</sup>.

Chatbots hold promise in increasing the efficiency of outpatient care pathways and meeting the need for patient surveillance and education between face-to-face clinic appointments. Accuracy of information and patient safety, however, are important considerations. The aim of this systematic review was to determine the uptake, acceptability and utility of chatbots in the follow-up of patients who have received physical healthcare interventions.

## Methods

The systematic review was designed and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement<sup>8</sup>. The protocol was

Received: March 24, 2021. Accepted: June 17, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of BJS Society Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

prospectively registered in the PROSPERO database (registration number: CRD42020199919)<sup>9</sup>.

## Search strategy

Search strategies included free text and index terms related to the following core concepts: 'chatbot' 'intervention' and 'follow-up' (Fig. S1, supplementary material). The following databases were searched from inception until 18 September 2020: MEDLINE, MEDLINE In Process, EMBASE, Cochrane CENTRAL, CINAHL and PsychINFO. The Central database was searched for registered clinical trials up until 9 November 2020. The search was not restricted by language or date of publication. A further search of the surgical grey literature was conducted by examining the proceedings of the 2020 Association of Surgeons in Training International Surgical Conference<sup>10,11</sup>.

## Eligibility criteria

All studies reporting original data were eligible for inclusion, including randomised trials, quasi-experimental designs, cohort studies, case-control studies and case series. Case reports, reviews, meta-analyses and articles related to the technical development of systems without accompanying clinical data were excluded. Systematic reviews were screened for potentially eligible publications. The titles and abstracts of identified articles were independently screened by two authors.

## Participants

Adult and paediatric patients who had undergone any physical healthcare intervention targeting physical rather than mental health and who were subsequently followed up using an automated conversational agent (a chatbot) at any point after an intervention were eligible for inclusion. Physical interventions were defined as procedures where purposeful access to the body was gained via an incision, percutaneous puncture or instrumentation via a natural orifice or the provision of medications to treat underlying disease. Examples of physical interventions included total hip replacement for osteoarthritis, steroid injection for carpal tunnel syndrome, transurethral resection of the prostate for benign prostatic hyperplasia and the prescription of antihypertensive medication.

## Interventions and comparators

A chatbot was defined as a computer software application that permits two-way conversation (via text, speech or a combination of both) between a human user and a computer program<sup>3</sup>. Comparators included other automated or non-automated follow-up systems, including, for example, routine care delivered via face-to-face outpatient clinics and follow-up telephone calls.

## Outcomes

The primary outcome assessed was the acceptability of chatbots as a method of follow-up indicated by implementation success. Measures of acceptability included user engagement (defined as the proportion of patients who activated and interacted with the chatbot), patient adherence to the chatbot, response rate (defined as the proportion of patients responding to system queries), duration of adherence and interactions with the chatbot over time. Patient safety and accuracy statistics were assessed where reported. Additional outcomes assessed included patient cohort demographics, design features such as task orientation, dialogue management, input and output formats, platforms used, health questionnaires used and measures of patient satisfaction.

## Study selection

Potentially eligible studies were compiled, and duplicate citations removed. Two authors independently screened titles and abstracts of retrieved studies using prespecified stepwise inclusion/exclusion criteria. Disagreements between reviewers were resolved through consultation with a third reviewer. Reference lists of included studies and published narrative/systematic reviews were examined for further potentially eligible studies.

## Data extraction and analysis

Data were extracted using a predefined electronic data-collection form. Extracted data were collated, cross-checked by other authors and compared. Study setting, population demographics, healthcare interventions, cohort-specific factors, software design features, measures of adherence, patient experience and clinical outcomes were extracted. Formal meta-analysis was not performed due to heterogeneous outcome reporting and differences in study designs. A narrative synthesis and descriptive analysis were used.

## Risk of bias analysis

Methodological quality of each included study was assessed. For randomised trials, this involved the revised Cochrane Risk of Bias tool<sup>12</sup>, and for non-randomised comparative studies the Cochrane Risk of Bias in non-randomised studies of interventions (ROBINS-I) tool<sup>13</sup>. The National Institute of Health (NIH) quality assessment tool for cohort studies was employed to assess the quality of cohort studies<sup>14</sup>.

## Results

From a total of 908 potential studies, 709 remained for screening after removal of duplicates, of which 11 articles were finally assessed with 10 meeting full inclusion criteria (Fig. 1).

## Study characteristics

Three randomised control trials (RCTs) were identified. One involving 76 participants compared an automated text-based chatbot with standard postoperative care following upper or lower extremity fracture<sup>15</sup>. The second, involving 142 participants, compared an automated chatbot versus physician-generated advice for women who had undergone breast cancer treatment<sup>16</sup> and the third, with 45 participants, compared immediate versus delayed access to a chatbot in young patients affected by various cancers<sup>17</sup>.

The non-randomised comparative clinical study included 270 participants and compared an automated speech-based chatbot to manual telephone follow-up for patients who had undergone orthopaedic surgery<sup>18</sup>. The remaining six studies<sup>19–24</sup> were cohort studies based on an established definition<sup>25</sup>. Collectively, eight out of 10 studies were published between 2019 to 2020.

## Demographics

Of the 10 included studies, nine recruited adults, and one adolescents with a mean age of 15 years<sup>19</sup> (Table 1) resulting in a total of 5492 patients. Chatbots were used to follow up patients after elective orthopaedic surgery<sup>18</sup>, orthopaedic trauma surgery<sup>15</sup>, surgical intervention for varicose veins<sup>21</sup>, women treated for breast cancer<sup>16,24</sup>, ureteroscopy<sup>23</sup> as well as the medical management of hypertension<sup>20</sup>, asthma<sup>19</sup> and various cancers<sup>17,22</sup>.

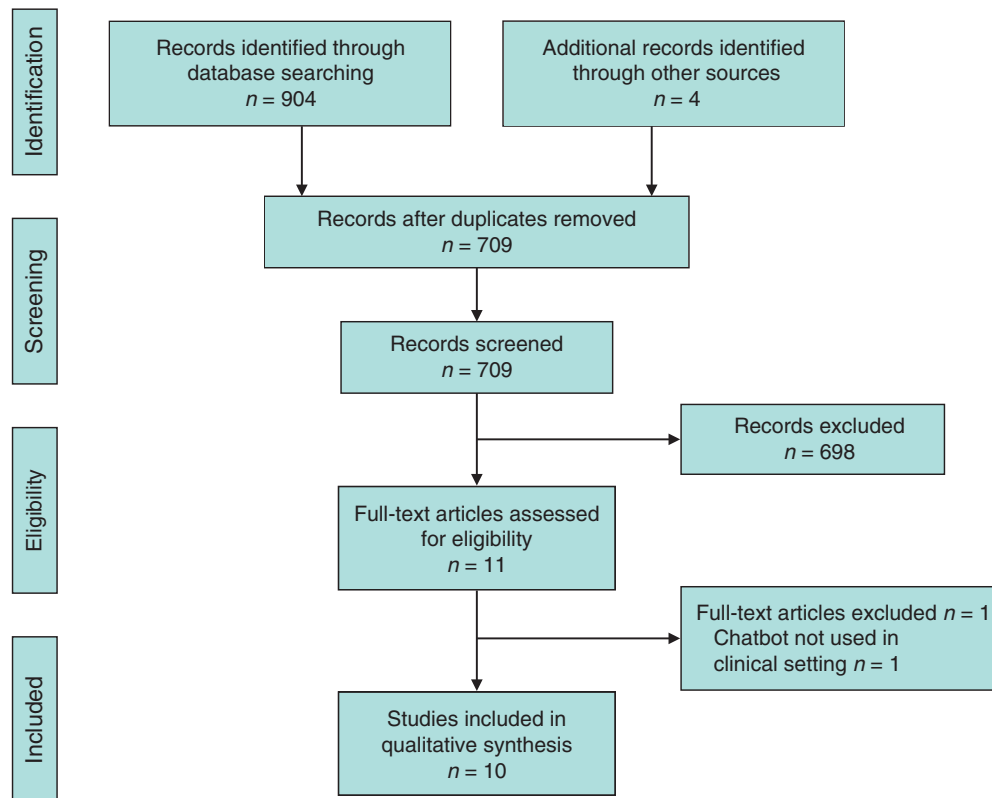


Fig. 1 PRISMA flow diagram

## Quality of included studies

One RCT was deemed to have a high risk of bias due to ascertainment bias and risk of detection bias given the effect of unblinding on the outcome of interest<sup>17</sup>. The remaining two RCTs were deemed at moderate risk of conduct bias<sup>15,16</sup>.

The cohort studies were rated as fair<sup>19,22,23</sup> or poor quality<sup>20</sup> (Fig. S2a-c, supplementary material). The quality of outcome measurement and assessment was deemed poor across all cohort studies.

## Interventions

All studies deployed chatbots on mobile devices: two were also accessible via web-based applications<sup>16,24</sup> and one was accessible via Facebook Messenger<sup>17</sup>. In terms of chatbot construct, seven used a frame-based knowledge-representation system<sup>16-20,23,24</sup>, one used a rule-based knowledge-representation system<sup>22</sup> and two studies did not report the type of system used<sup>15,21</sup>. Of the 10 studies, three used a system-focused dialogue<sup>15,22,23</sup>, two a user-focused dialogue<sup>16,24</sup> and the other five used a mixed dialogue initiative<sup>17-21</sup>. Task orientation was reported in two studies, one chatbot was able to book follow-up appointments<sup>21</sup> and one was able to input patient data into electronic medical records<sup>20</sup>.

## Outcomes

Measures of implementation success were reported in seven of 10 studies<sup>17-19,21-24</sup>. Adherence ranged from 31 per cent participant engagement rate<sup>24</sup> to 97 per cent participant response rate for select system-generated questions<sup>19</sup>. One study demonstrated a decline in engagement from 100 to 31 per cent after 8 months of chatbot use<sup>24</sup>. A comparative study demonstrated a 92 per

cent follow-up rate for patients contacted via an autonomous postoperative chatbot versus a 93 per cent follow-up rate for patients contacted directly by phone<sup>18</sup>.

Other outcome measures reported by studies included patient-reported outcome scores (PROMs), patient feedback, patient experience and technical details related to chatbot performance (Table 2). One RCT demonstrated that a chatbot with twice-daily text-based output for 2 weeks was associated with reduced opiate consumption compared with a control cohort (no messages received) following orthopaedic trauma surgery (26 opiate tablets versus 41 tablets)<sup>15</sup>. Another RCT found no differences in perceived quality of responses using the between chatbot versus real-time physician-written responses to user queries from women treated for breast cancer (average QLQ-INFO25 score 2.89 and 2.82 respectively)<sup>16</sup>. The third RCT reported no significant difference in symptoms of anxiety and depression, quantified using the Emotional Disturbance Anxiety Score, between patients using a chatbot (cohort 1) and a control cohort without chatbot access (cohort 2) over a 4-week study period. Upon completion of the first study period, the control cohort (cohort 2) were then granted access to the chatbot and symptoms of anxiety and depression were quantified after a second 4-week study period. After the second study period, patients in cohort 2 demonstrated a reduction in reported symptoms of anxiety compared with baseline measurements and anxiety scores after the first study period, although this reduction was not statistically significant<sup>17</sup>. A non-randomised comparative study demonstrated comparable follow-up consultation rates after orthopaedic surgery using a telephone-based conversational agent compared with calls made by individuals, saving an estimated 9.3 hours per 100 participants<sup>18</sup>.

Table 1 Study demographics, quality and risk of bias

Reference	Study type	n	Speciality	Cohort	Intervention (+ control)	Study quality	Risk of bias
<b>Anthony et al.</b> , 2020 <sup>15</sup>	RCT	76	Orthopaedics	Adult patients who had undergone operative fixation of upper or lower extremity fracture	Automated chatbot which delivered text messages to reduce opioid use versus standard postoperative care	–	Some concerns*
<b>Bian et al.</b> , 2020 <sup>18</sup>	Comparative clinical study	270	Orthopaedics	Adult patients who had undergone orthopaedic intervention	Automated chatbot versus telephone follow-up for postoperative care	–	Moderate†
<b>Bibault et al.</b> , 2019 <sup>16</sup>	RCT	142	Oncology	Adult patients in remission or undergoing active treatment for breast cancer	Automated chatbot which provides information to patients about breast cancer, epidemiology, treatment options, side effects and quality of life improvement strategies versus information provided by treating physicians in real time via text message	–	Some concerns*
<b>Black et al.</b> , 2020 <sup>21†</sup>	Prospective cohort	158	Vascular surgery	Adult patients undergoing intervention for lower extremity superficial venous reflux with endovascular ablation, sclerotherapy and phlebectomy	Automated postoperative chatbot offered to patients to educate patients, provide postoperative instructions, facilitate follow-up appointment booking and contact with the clinic	–	–
<b>Chaix et al.</b> , 2019 <sup>24</sup>	Prospective cohort	4737	Oncology	Adult patients in remission or undergoing active treatment for breast cancer	Automated chatbot which provides information to patients about breast cancer, epidemiology, treatment options, side effects and quality-of-life improvement strategies	Fair	–
<b>Giorgino et al.</b> , 2005 <sup>20</sup>	Prospective cohort	15	Internal medicine	Adult patients with diagnosed hypertension treated with oral medications	Automated chatbot used to monitor hypertensive patients in the community. Collects health-related data such as heart rate and blood pressure	Poor	–
<b>Goldenthal et al.</b> , 2019 <sup>23</sup>	Prospective cohort	20	Urology	Adult patients who had undergone ureteroscopy for nephrolithiasis within the previous month	Automated chatbot used to educate and reassure patients regarding commonly experienced symptoms or post-procedural complications	Fair	–
<b>Greer et al.</b> , 2019 <sup>17</sup>	RCT	45	Oncology	Young adult patients (aged 18–25 years) who had completed active treatment for cancer within the past 5 years	Automated chatbot used to provide cognitive and behavioural intervention that develops eight positive psychological skills. Patients were given conversational teaching sessions and practice lessons. Control participants were asked to provide daily emotion ratings. The control group had no access to the chatbot but were given access after 4 weeks	–	High*
<b>Piau et al.</b> , 2019 <sup>22</sup>	Prospective cohort	9	Internal medicine	Adult patients aged >65 years with a diagnosis of cancer undergoing active treatment with chemotherapy	Automated chatbot used to identify the development of symptoms or treatment side effects	Fair	–
<b>Rhee et al.</b> , 2014 <sup>19</sup>	Prospective cohort	15	Internal medicine	Adolescent patients (and patient dyads) diagnosed with asthma receiving active treatment	Automated chatbot used to monitor patient symptoms, activity levels and medication use	Fair	–

\* As per RoB 2 tool. † As per ROBINS-1 tool. ‡ Quality appraisal and risk of bias assessment not performed as full manuscript not published (data extracted from conference proceedings).

Table 2 Technical details, acceptability criteria and outcomes assessed

Reference	Chatbot features	Device	Adherence	Other outcomes measured
<b>Anthony et al., 2020<sup>15</sup></b>	<ul style="list-style-type: none"> <li>System-focused dialogue initiative</li> <li>Text output</li> </ul>	Smartphone (text)	Not reported	<p><b>Postoperative pain</b></p> <ul style="list-style-type: none"> <li>36.5% reduction in number of opiate tablets used in intervention group (<math>P = 0.004</math>)</li> <li>35% decrease in morphine milliequivalents consumed <i>versus</i> control (<math>P = 0.006</math>)</li> </ul> <p><b>Patient-reported outcomes</b></p> <ul style="list-style-type: none"> <li>Lower mean postoperative pain intensity 3 A PROMIS score in intervention arm (<math>45.9 \pm 7.2</math> <i>versus</i> <math>49.7 \pm 8.8</math>, <math>P = 0.04</math>)</li> <li>Lower mean postoperative pain interference 8 A PROMIS score in intervention arm (<math>60.6 \pm 8.2</math> <i>versus</i> <math>56.6 \pm 9.4</math>, <math>P = 0.04</math>)</li> </ul> <p><b>Patient feedback</b></p> <ul style="list-style-type: none"> <li>10.3% of patients contacted via chatbot provided feedback (<i>versus</i> 2.5% control)</li> </ul> <p><b>Time per 100 patients</b></p> <ul style="list-style-type: none"> <li>0 <i>versus</i> 9.3 hours for chatbot and control respectively</li> </ul> <p><b>Quality of response</b></p> <ul style="list-style-type: none"> <li>Perceived quality of response to the answers provided to user queries assessed using the QLQ-INFO25 (a patient-satisfaction score). Patients assessing chatbot responses gave a higher average rating compared with rating for responses given by physicians in real time. Success was defined as a score greater than or equal to 3 on a satisfaction scale of 1–4. Overall, non-inferiority was demonstrated between perceived quality of responses, however when individual items of the QLQ-INFO25 were assessed individually, non-inferiority of response satisfaction could not be demonstrated in 9 of 25 items</li> </ul>
<b>Bian et al., 2020<sup>18</sup></b>	<p>Frame based</p> <p>Mixed dialogue initiative</p> <p>Spoken input/output</p>	Smartphone (call)	92.2% follow-up rate ( <i>versus</i> 93.3% in control)	
<b>Bibault et al., 2019<sup>16</sup></b>	<ul style="list-style-type: none"> <li>Frame based</li> <li>User-focused dialogue</li> <li>Text input and output</li> </ul>	Web based or smartphone application	Not reported	
<b>Black et al., 2020<sup>21</sup></b>	<ul style="list-style-type: none"> <li>Mixed dialogue initiative</li> <li>Orientated to book follow-up appointments</li> </ul>	Smartphone (application)	83.3% of participants engaged with the chatbot	<p><b>Patient satisfaction</b></p> <ul style="list-style-type: none"> <li>59% of patients wanted more information (<i>versus</i> 65% control)</li> <li>85% of patients found information useful (<i>versus</i> 83.1% control)</li> <li>85% of patients satisfied with amount of information received (<i>versus</i> 77% control)</li> </ul> <p><b>Patient experience</b></p> <ul style="list-style-type: none"> <li>60% highly satisfied (rated chatbot useful or very useful)</li> </ul>
<b>Chaix et al., 2019<sup>24</sup></b>	<ul style="list-style-type: none"> <li>Frame based</li> <li>User-focused dialogue</li> <li>Text input and output</li> </ul>	Web based or smartphone application	31% retention rate after 8 months (N.B. only calculated for 956 patients)	<p><b>User response characteristics</b></p> <p>Average response length 21.5 words</p> <p><b>Patient experience</b></p> <ul style="list-style-type: none"> <li>93.95% overall patient satisfaction</li> <li>88% stated that chatbot provided them with support and helped them follow their treatment effectively</li> </ul> <p><b>Technical details</b></p> <ul style="list-style-type: none"> <li>80% consultation conclusion rate reached by system</li> <li>18 questions per interaction</li> <li>Average consultation call time 3.3 minutes</li> </ul>
<b>Giorgino et al., 2005<sup>20</sup></b>	<ul style="list-style-type: none"> <li>Frame based</li> <li>Mixed dialogue initiative</li> <li>Spoken input/output</li> <li>Orientated to integrate with medical records</li> </ul>	Smartphone (call)	Not reported	
<b>Goldenthal et al., 2019<sup>23</sup></b>	<ul style="list-style-type: none"> <li>Frame based</li> </ul>	Smartphone (application)	35% of participants engaged with chatbot	<p><b>Reasons for not activating chatbot</b></p>

(continued)

Table 2. (continued)

Reference	Chatbot features	Device	Adherence	Other outcomes measured
Greer et al., 2019 <sup>17</sup>	<ul style="list-style-type: none"> <li>System-focused dialogue initiative</li> <li>Text input/output</li> </ul> Frame based Mixed dialogue initiative Text input and output	Smartphone (Facebook messenger)	Mean of 12.1 sessions (73.8 minutes total engagement time) across 4 weeks versus 18.1 sessions (27.1 minutes total engagement time)	Misplacing instructions for chatbot use ( $n = 6$ ), relying on follow-up with clinic or discharge materials ( $n = 4$ ), inability to activate chatbot ( $n = 2$ ) and inability to text ( $n = 1$ )  <b>Patient satisfaction</b> Patients rated chatbot as useful (average score 2/3) Patients likely to recommend chatbot to friend (average rating 6.9/10) <b>Anxiety and depression symptoms</b> Participants in intervention arm reported greater reduction in anxiety versus the control arm as per the PROMIS Emotional Distress-Anxiety Short Form (2.58 t-score units versus 0.7, $P = 0.09$ ) Both intervention and control arms reported a reduction in depressive symptoms as per the PROMIS Emotional Distress-Depression Short Form (1.83 versus 1.38, $P = 0.77$ )
Piau et al., 2019 <sup>22</sup>	<ul style="list-style-type: none"> <li>Rule based</li> <li>System-focused dialogue initiative</li> <li>Voice and text input/output</li> </ul>	Smartphone (application)	86% compliance over study period	<b>Patient experience</b> 3 participants provided feedback (33%) All three were satisfied or very satisfied <b>Technical details</b> 3.5-minute average time to complete consultation All patients had a smartphone prior to recruitment
Rhee et al., 2014 <sup>19</sup>	Frame based Mixed dialogue initiative Text input/output	Smartphone (application)	81–97% response rate for system-initiated questions	<b>Patient experience</b> High overall satisfaction reported <b>Technical details</b> Average number of user initiated questions: 19

PROMIS, Patient Reported Outcomes Measurement Information System.

## Registered trials

The authors' search found two additional registered protocols for ongoing clinical trials. Study protocols outline the intended use of chatbots to facilitate questionnaire completion at 6 and 8 months following bariatric surgery<sup>26</sup> and for daily consultation with patients treated for Parkinson's disease<sup>27</sup> (Fig. S3, supplementary material).

## Discussion

The use of chatbots following a physical healthcare intervention is a new and evolving field, with eight of 10 studies published during or after 2019. It seems likely that this will continue to increase, with a move towards efficiency in healthcare systems and a move away from face-to-face follow-up arising from the COVID-19 pandemic.

A review investigating the broader use of conversational agents in healthcare has been published<sup>4</sup>, while the present review was focused on the role of technology after interventions. The systematic review identified 10 studies of different designs, mostly of moderate to poor quality. All outcome measures were inconsistently defined and outcome assessors were not blinded, predisposing to detection bias and Hawthorne effect. One study attempted to reduce this by blinding participants to responses from either the chatbot or physicians<sup>16</sup>, although by the nature of the intervention, a Hawthorne effect cannot be ignored.

Acceptability and patient experience using automated conversational agents was largely positive<sup>19,21</sup>. There was no clinically important difference in rates of patient satisfaction with chatbot responses compared with real-time physician-generated responses to user queries, measured using the QLQ-INFO25<sup>16</sup>. Previous work has demonstrated the QLQ-INFO25 is acceptable with good internal consistency and test-retest reliability<sup>28</sup>. The reduction in opiate prescribing, time and cost saving reported in one small study provides useful evidence supporting investment in automated follow-up systems<sup>15</sup>.

Despite the metrics used being heterogeneous, data around success of implementation suggest considerable variation. Some learning points were simple and applicable. One study described a 35 per cent interaction rate with their chatbot, with the primary reason for poor interaction being 'misplacing instructions for chatbot use'<sup>23</sup>, while another demonstrated an initial engagement rate of 100 per cent at the start of the study that gradually fell to 31 per cent over 8 months<sup>24</sup>, likely to represent reduced enthusiasm for patient engagement, although it might represent patient adaptation to their current health state. Some support for the latter is that most (88 per cent) participants reported that the chatbot provided them with support and helped them follow their treatment plan. A structured sequence to implementation may increase success, and frameworks for this have been developed for the deployment of PROMs that might be applicable to automated follow-up systems<sup>29</sup>.

No study identified in the current systematic review examined patient safety. If autonomous agents are to be used in clinical practice to monitor patient status actively after intervention, rigorous safety testing using simulated patients is warranted before clinical adoption. Following implementation, prospective registries of technological adverse events should be kept. Here, technological adverse events refer to patient harm directly caused by technology. This harm may be direct (inappropriate clinical advice) or indirect (failing to identify clinical signs of deterioration). All studies identified in this systematic review deployed agents

on mobile devices. In the UK, 70 per cent of adults own a smartphone and over half regularly use applications<sup>30</sup>. Disparities in socioeconomic status and technological literacy may limit access to healthcare. Future epidemiological studies should seek to ascertain whether clinical implementation of technologies negatively impacts the health of certain cohorts within the population.

The present study has a number of limitations. A small number of heterogeneous studies were identified, reporting a variety of different adherence and clinical-outcome measures. The majority of studies were small, non-comparative feasibility studies. The comparative studies were at risk of selection and detection bias owing to the nature of interventions and relative infancy of the field. Varying technical descriptions of agents were provided and heterogeneity in outcome reporting precluded meaningful meta-analysis, limiting the strength of conclusions that can safely be drawn.

There is, nevertheless, early evidence of uptake of automated conversational agents in the outpatient management of patients following physical healthcare interventions. Despite a range of chatbot builds and clinical uses, they seem to be generally acceptable, although effectiveness remains to be proven. Attention to practical details around deployment may improve implementation success of future systems.

## Acknowledgements

L.G. was involved in idea inception, search strategy design, data extraction, analysis and writing. A.S. and J.C.R.W. were involved in abstract screening and manuscript review. C.J.H., D.C. and M.G. critically reviewed the manuscript. J.B. and J.N.R. were involved in idea inception, search strategy design and manuscript review.

## Funding

No specific funding was received for the conduct of this review. J.B. is supported by National Institute for Health Research Capability Funding via University Hospitals Coventry and Warwickshire. C.J.H. is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684). J.N.R. is funded by an NIHR Postdoctoral Fellowship (PDF-2017-10-075). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Disclosure. The authors declare no conflicts of interest.

## Supplementary material

Supplementary material is available at *BJS Open* online

## References

- Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun ACM* 1966;**9**:36–45
- Adamopoulou E, Moussiades L. An overview of chatbot technology. *Artif Intell Appl Innov* 2020;**584**:373–383
- Lexico. Chatbot. <https://www.lexico.com/definition/chatbot> (accessed 7 December 2020)
- Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc* 2018;**25**:1248–1258
- Parkes RJ, Palmer J, Wingham J, Williams DH. Is virtual clinic follow-up of hip and knee joint replacement acceptable to patients and clinicians? A sequential mixed methods evaluation. *BMJ Open Qual* 2019;**8**:e000502
- Scarborough A, Geoghegan L, Horwitz MD, Naqui Z. Implementation of virtual consultation for hand surgeons and therapists: an international survey and future implications. *J Hand Surg Eur Vol* 2020;**45**:1005–1006
- Healy P, McCrone L, Tully R, Flannery E, Flynn A, Cahir C et al. Virtual outpatient clinic as an alternative to an actual clinic visit after surgical discharge: a randomised controlled trial. *BMJ Qual Saf* 2019;**28**:24–31
- Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;**6**:e1000100
- Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Syst Rev* 2012;**1**:2
- ASIT oral presentations. *Brit J Surg* 2020;**107**:5–24
- ASIT poster presentations. *Brit J Surg* 2020;**107**:25–234
- Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;**366**:l4898
- Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;**355**:i4919
- National Institutes of Health. Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies. <https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools> (accessed 24 March 2021)
- Anthony CA, Rojas EO, Keffala V, Glass NA, Shah AS, Miller BJ, Hogue M et al. Acceptance and commitment therapy delivered via a mobile phone messaging robot to decrease postoperative opioid use in patients with orthopedic trauma: randomized controlled trial. *J Med Internet Res* 2020;**22**:e17750
- Bibault J-E, Chaix B, Guillemassé A, Cousin S, Escande A, Perrin M et al. A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled non-inferiority trial. *J Med Internet Res* 2019;**21**:e15787
- Greer S, Ramo D, Chang Y-J, Fu M, Moskowitz J, Haritatos J. Use of the chatbot ‘vivibot’ to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. *JMIR mHealth Uhealth* 2019;**7**:e15018
- Bian Y, Xiang Y, Tong B, Feng B, Weng X. Artificial intelligence-assisted system in postoperative follow-up of orthopedic patients: exploratory quantitative and qualitative study. *J Med Internet Res* 2020;**22**:e16896
- Rhee H, Allen J, Mammen J, Swift M. Mobile phone-based asthma self-management aid for adolescents (mASMAA): a feasibility study. *Patient Prefer Adherence* 2014;**8**:63–72
- Giorgino T, Azzini I, Rognoni C, Quaglini S, Stefanelli M, Gretter R et al. Automated spoken dialogue system for hypertensive patient home management. *Int J Med Inform* 2005;**74**:159–167
- Black C, Nelson B, DeShazo J. 3:18 PM Abstract No. 74 Feasibility of implementing medical chatbot technology into the postoperative care for patients undergoing treatment for superficial venous insufficiency. *J Vas Interv Radiol* 2020;**31**:S36

22. Piau A, Crissey R, Brechemier D, Balardy L, Nourhashemi F. A smartphone chatbot application to optimize monitoring of older patients with cancer. *Int J Med Inform* 2019;**128**:18–23
23. Goldenthal SB, Portney D, Steppe E, Ghani K, Ellimoottil C. Assessing the feasibility of a chatbot after ureteroscopy. *Mhealth* 2019;**5**:8
24. Chaix B, Bibault J-E, Pienkowski A, Delamon G, Guillemassé A, Nectoux P *et al*. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR Cancer* 2019;**5**:e12856
25. Dekkers OM, Egger M, Altman DG, Vandenbroucke JP. Distinguishing case series from cohort studies. *Ann Intern Med* 2012;**156**:37–40
26. Follow-up of Bariatric Surgery by Teleconsultations. <https://www.cochranelibrary.com/central/doi/10.1002/central/CN-02181004/full> (accessed 7 December 2020)
27. A Study of AI Conversation in Telemedicine for Parkinson's Disease. <https://www.cochranelibrary.com/central/doi/10.1002/central/CN-02172801/full> (accessed 7 December 2020)
28. Arraras JI, Greimel E, Sezer O, Chie WC, Bergenmar M, Costantini A *et al*. An international validation study of the EORTC QLQ-INFO25 questionnaire: an instrument to assess the information given to cancer patients. *Eur J Cancer* 2010;**46**:2726–2738
29. Stover A, Haverman L, van Oers HA, Greenhalgh J, Potter CM. Using an implementation science approach to implement and evaluate patient-reported outcome measures (PROM) initiatives in routine care settings. *Qual Life Res* 2020. <https://link.springer.com/article/10.1007%2Fs11136-020-02564-9> (accessed 24 March 2021)
30. Ofcom: Adults Media Use and Attitudes Report 2016. [https://www.ofcom.org.uk/\\_\\_data/assets/pdf\\_file/0026/80828/2016-adults-media-use-and-attitudes.pdf](https://www.ofcom.org.uk/__data/assets/pdf_file/0026/80828/2016-adults-media-use-and-attitudes.pdf) (accessed 4 November 2020)